A PROJECT REPORT ON

"CHRONIC KIDNEY DISEASE PREDICTION SYSTEM"

(Submitted in fulfilment of SmartBridge-Smartinternz- RSIP-2020)

Smartbridge Trading Solutions Pvt Ltd

SUBMITTED TO: MS. NIDHI MAM

SUBMITTED BY: PRATEEK GUPTA, **B.TECH - CSE, Dr.** Akhilesh Das Gupta Institute of Technology and Management, **ADGITM- GGSIPU , DELHI**

AKULA NAVEEN, , **B.TECH – ECE, Sri Venkateshwara College of Engineering and Technology, Etcherla, Srikakulam, West Bengal**

LAMMATHA SUNEETHA , **B.TECH – ECE, , Sri Venkateshwara College of Engineering and Technology, Etcherla, Srikakulam, Andhra Pradesh**

DATE: 9th June 2020 to 7th July 2020

❖ DECLARATION

We here by declare that the work done on the dissertation entitled " Chronic Kidney Disease Prediction System " has been carried out by us and submitted in the fulfilment of the **SmartBridge- Smartinternz- RSIP-2020** in '***Machine Learning with Python***' at **SmartBridge**, India

❖ ACKNOWLEDGEMENT

The satisfaction that accompanies the successful completion of any task would be incomplete without the mention of people whose cooperation made it possible, whose constant guidance and encouragement crown all the efforts with success.

We all grateful to our project mentor Ms. **Nidhi Mam**, **Vinay Sir** for the guidance, inspiration and constructive suggestion that helped us in the preparation of the project.

And finally, We extend our heartfelt gratitude to our parents and friends whose constant support helped me to emerge with a successful project.

❖ TABLE OF CONTENT

8	**RESULT & DISCUSSSION**
9	**CONCLUSON** • Scope for further development • Scope for enhancement
10	• Bibliography ➢ References ➢ Websites ➢ Books

ABSTRACT

Chronic Kidney Disease (CKD) is one of the most widespread illnesses in the United States. Recent statistics show that twenty-six million adults in the United States have CKD and million others are at increased risk. Clinical diagnosis of CKD is based on blood and urine tests as well as removing a sample of kidney tissue for testing. Early diagnosis and detection of kidney disease is important to help stop the progression to kidney failure. Data mining and analytics techniques can be used for predicting CKD by utilizing historical patient's data and diagnosis records. In this research, predictive analytics techniques such as Decision Trees, Logistic Regression, Naive Bayes, and Artificial Neural Networks are used for predicting CKD. Pre-processing of the data is performed to impute any missing data and identify the variables that should be considered in the prediction models. The different predictive analytics models are assessed and compared based on accuracy of prediction. The study provides a decision support tool that can help in the diagnosis of CKD.

Chronic Kidney Disease prediction is one of the most important issues in healthcare analytics. The most interesting and challenging tasks in day to day life is prediction in medical field. In this paper, we employ some machine learning techniques for predicting the chronic kidney disease using clinical data. We use three machine learning algorithms such as Decision Tree(DT) algorithm, Naive Bayesian (NB) algorithm. The performance of the above models are compared with each other in order to select the best classifier in predicting the chronic kidney disease for given dataset. Ultimately, gradient boosted decision trees proved to be the best prediction model with a predictive accuracy of 83.94%.

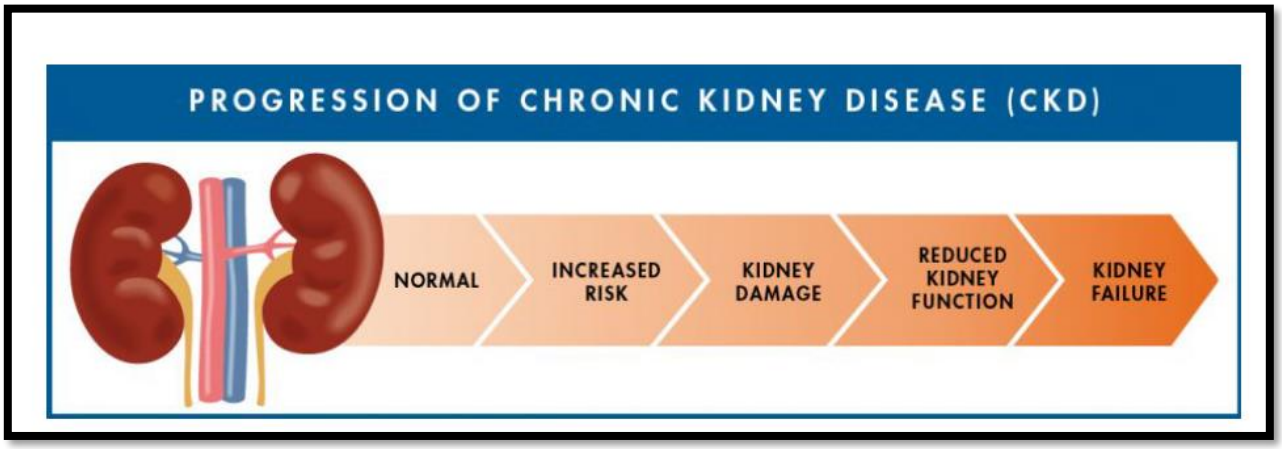
Chronic kidney disease (CKD), also known as chronic renal disease. Chronic kidney disease involves conditions that damage your kidneys

and decrease their ability to keep you healthy. You may develop complications like high blood pressure, anemia (low blood count), weak bones, poor nutritional health and nerve damage. . Early detection and treatment can often keep chronic kidney disease from getting worse. Data Mining is the term used for knowledge discovery from large databases. The task of data mining is to make use of historical data, to discover regular patterns and improve future decisions, follows from the convergence of several recent trends: the lessening cost of large data storage devices and the ever-increasing ease of collecting data over networks; the expansion of robust and efficient machine learning algorithms to process this data; and the lessening cost of computational power, enabling use of computationally intensive methods for data analysis. Machine learning, has already created practical applications in such areas as analysing medical science outcomes, detecting fraud, detecting fake users etc. Various data mining classification approaches and machine learning algorithms are applied for prediction of chronic diseases. The objective of this research work is to introduce a new decision support system to predict chronic kidney disease. The aim of this work is to compare the performance of Support vector machine (SVM) and K-Nearest Neighbour (KNN) classifier on the basis of its accuracy, precision and execution time for CKD prediction. From the experimental results it is observed that the performance of KNN classifier is better than SVM.

INTRODUCTION

The Problem

Machine learning has been used to detect CKD and its progression using labwork data. While this is potentially useful for those patients that find themselves getting bloodwork done at routine checkups or in tandem with other health-related issues, it poses two problems: first, it requires individual patients to have regular lab work done, and second, it does nothing to predict the disease at the population level. The Centers for Disease Control (CDC) recognize chronic kidney disease as a public health concern that requires population-level surveillance and prevention. However, the hyperlocal surveillance of CKD is difficult and costly. This limits our ability to effectively target campaigns to prevent CKD and its progression. In order to predict hyperlocal disease prevalence, the model presented in this paper uses readily available Census Bureau and CDC data on known and hypothesized risk factors with stochastic gradient descent to better identify census tracts where aggressive public health campaigns and healthcare initiatives can positively affect the early detection and treatment of CKD. Further investigation into this model may also give insight into potential risk factors.



Chronic Kidney Disease (CKD) is a progressive condition that results in significant morbidity and mortality. Because of the important role the kidneys play in maintaining homeostasis, CKD can affect almost every body system. Early recognition and intervention are essential to slowing disease progression maintaining quality of life, and improving outcomes. Family physicians have the opportunity to screen at-risk patients, identify affected patients, and ameliorate the impact of CKD by initiating early therapy and monitoring disease progression.

The purpose of this case is to create an easy-to-use screening tool to identify patients at risk for CKD. Despite the wide availability and low cost of a test for CKD based on one or more blood samples, studies have shown that many in the at-risk population have not been tested. One reason for this is that awareness of CKD is low. Given the proven benefits of early detection and treatment, the need for some kind of screening tool is clear. Although there is no reason to test everyone, those patients with a high enough probability of having CKD should be tested. The purpose of this case is to see if those high-risk patients can be identified using easily obtainable patient data.

The Causes of CKD

The two main causes of chronic kidney disease are diabetes and high blood pressure, which are responsible for up to two-thirds of the cases. Diabetes happens when your blood sugar is too high, causing damage to many organs in your body, including the kidneys and heart, as well as blood vessels, nerves, and eyes. High blood pressure, or hypertension, occurs when the pressure of your blood against the walls of your blood vessels increases. If uncontrolled, or poorly controlled, high blood pressure can be a leading cause of heart attacks, strokes, and chronic kidney disease. Also, chronic kidney disease can cause high blood pressure.

Other conditions that affect the kidneys are:

• Glomerulonephritis, a group of diseases that cause inflammation and damage to the kidney's filtering units. These disorders are the third most common type of kidney disease.

• Inherited diseases, such as polycystic kidney disease, which causes large cysts to form in the kidneys and damage the surrounding tissue.

• Malformations that occur as a baby develops in its mother's womb. For example, a narrowing may occur that prevents normal outflow of urine and causes urine to flow back up to the kidney. This causes infections and may damage the kidneys.

• Lupus and other diseases that affect the body's immune system.

• Obstructions caused by problems like kidney stones, tumors, or an enlarged prostate gland in men.

• Repeated urinary infections.

The test used a formula to estimate glomerular filtration rate based on measured serum creatinine concentration, age, gender, and race. CKD was defined as estimated filtration rate less than 60 ml/min/1.73 m2. For details, see Heejung Bang, David A. Shoham, Philip J. Klemmer, Ronald

Who Is at Risk?

While anyone at any age can develop chronic kidney disease (CKD), a number of risks factors have been identified that may lead to possible problems with your kidneys. These include:

• **Diabetes.** Diabetes is the leading cause of CKD. If you have diabetes, talk with your doctor about how to keep your blood glucose as close to normal as possible to ensure your diabetes is under control.

• **Hypertension.** Hypertension, also called high blood pressure, is the second-highest cause of CKD. Keep your blood pressure under control. A number of effective medications are available to help you with this task. Your doctor will help you to determine which medication is right for you.

• **Cardiovascular disease.** In addition to hypertension, other diseases of the heart and blood vessels may increase your risk for kidney disease. People who have had heart attacks or strokes, congestive heart failure, coronary artery disease, or peripheral vascular disease need to be monitored carefully for kidney problems.

• **Family history of kidney disease.** Some kidney diseases are genetic. People with a mother, father, brother, or sister who has had a kidney disease are more likely to develop problems with their kidneys.

• **Age.** People 60 years and older are at a higher risk for developing CKD.

• **Race.** People belonging to certain ethnic groups, such as First Nations (Canadian aboriginal peoples) and Pacific Islanders, are at a higher risk for developing this disease.

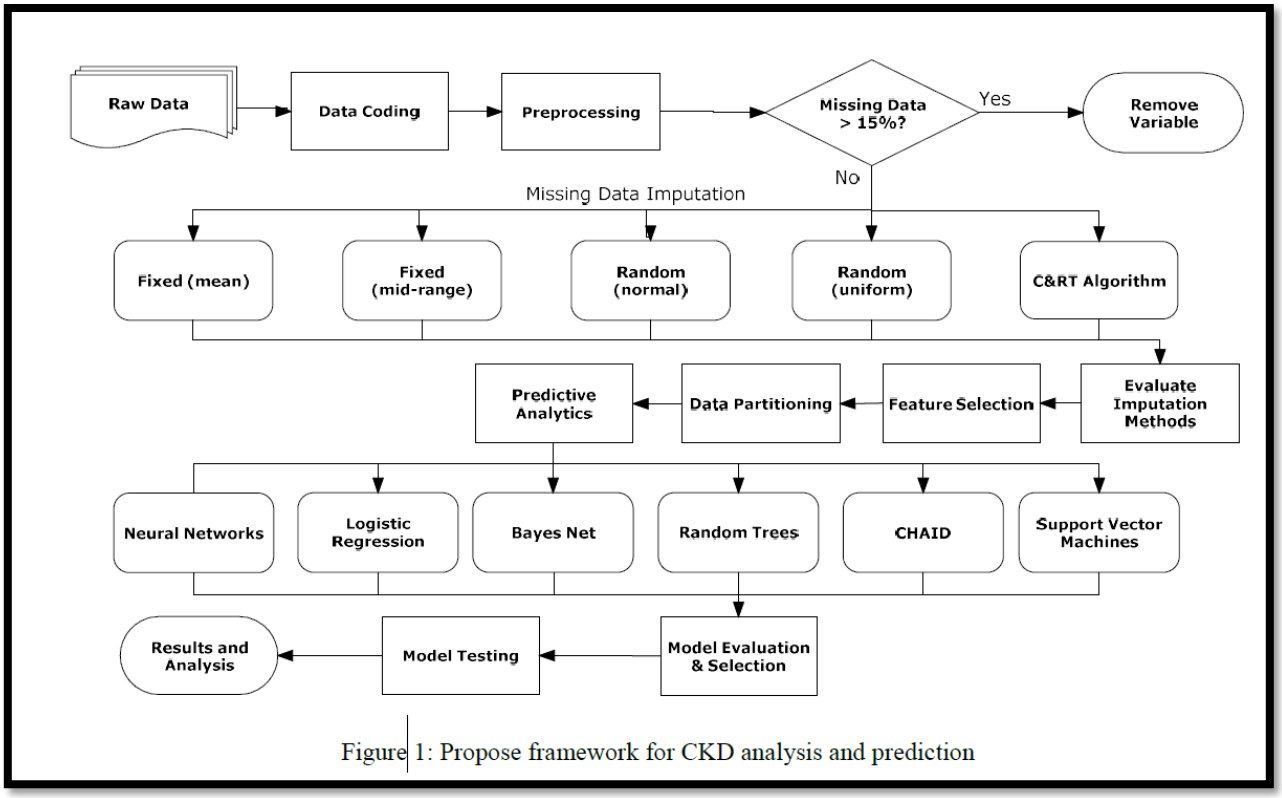


Figure 1: Propose framework for CKD analysis and prediction

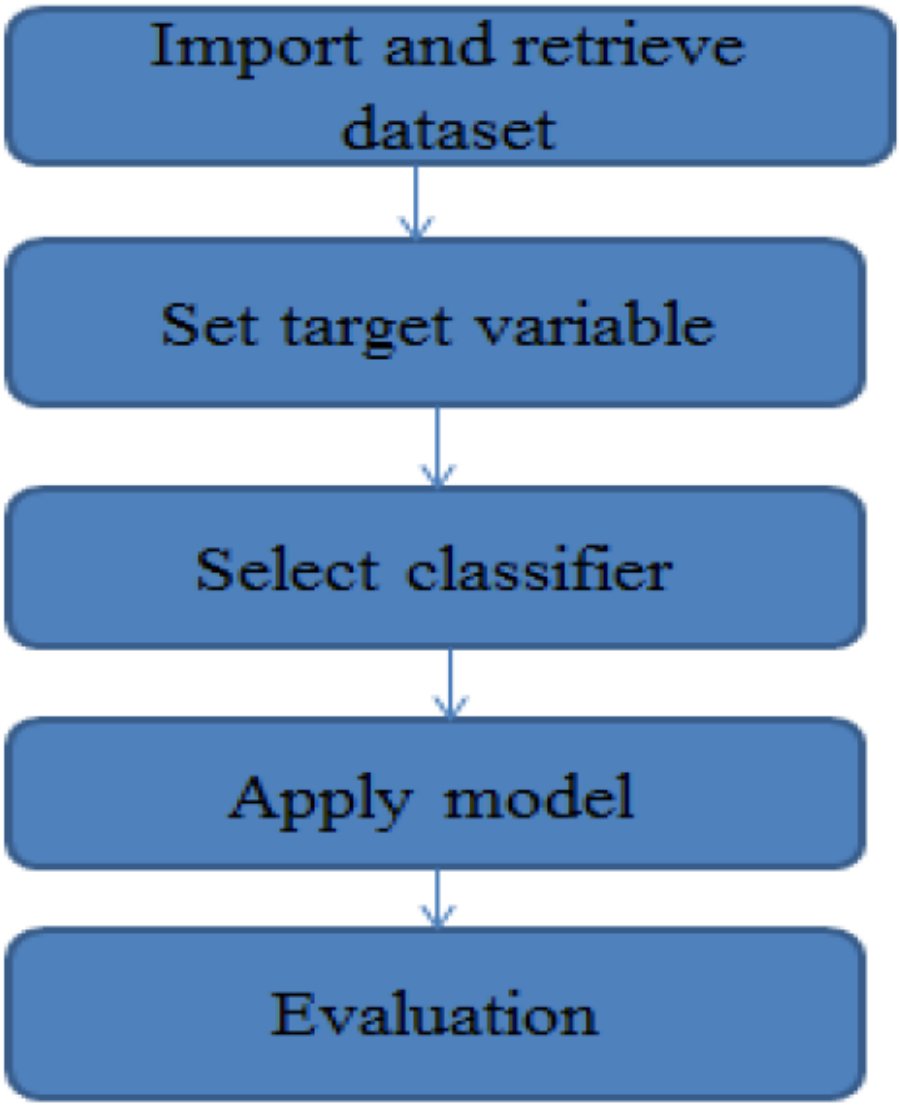


Fig 3.1 The process model to predict chronic kidney disease.

ABOUT THE PROJECT

Existing System

Chronic Kidney Disease (CKD) is a major medical problem and can be cured if treated it in the early stages. Usually, people are not aware that medical tests, we take for different purposes could contain valuable information concerning kidney diseases. Consequently, attributes of various medical tests are investigated to distinguish which attributes may contain helpful information about the disease. The information says that it helps us to measure the severity of the problem, the predicted survival of the patient after the illness, the pattern of the disease and work for curing the disease.

Proposed System

The work proposed here uses three classification techniques to predict the presence of chronic kidney disease in humans. The classifiers used are Support vector machine and KNN classifier. The data set for chronic kidney disease was gathered and applied on each classifier to predict the disease and the performance of the classifier is evaluated based on accuracy, precision and F measure. Architecture of Predictive Data Mining: Proposed Approach. The working of the architecture is as follows: The dataset for CKD patients have been collected and fed into the classifier named SVM and KNN. The prediction of CKD will be executed with the help of a tool known as MATLAB.

Disadvantages of system

The existing prediction system for chronic kidney disease is fine with some limitations. Below is the table shown, describing the worked done for prediction and detection of various kidney diseases. A new CKD prediction system is still the need. A decision support system for chronic kidney disease is still the need for early prediction, as not much work is done for the same. The list of risk factors above is a reflection of the results of several separate studies. What we want to do is figure out how to combine all the possible risk factors to measure the overall risk faced by the study subjects. The 34 variables in the data set are all easily obtained by a family physician during routine check-ups. Only the cholesterol measurements and the haemoglobin count.

Data Pre-Processing

The CDC data was paired down to include only tract-level data. It was then pivoted so that each row represents a census tract and each column a variable. Since the tract-level CDC data is not regularly updated, all disease variables other than chronic kidney disease were dropped from the dataset in order to avoid fitting the model to data that could not be easily attained beyond the 2015 data. This was done to create a model that could be better reproduced in the future with updated Census data. Data pre-processing is a process of cleaning the raw data i.e. the data is collected in the real world and is converted to a clean data set. In other words, whenever the data is gathered from different sources it is collected in a raw format and this data isn't feasible for the analysis.

Therefore, certain steps are executed to convert the data into a small clean data set, this part of the process is called as data pre-processing

Follow the following steps to process your Data

- Import the Libraries
- Importing the dataset
- Taking care of Missing Data
- Label encoding

- One Hot Encoding
- Feature Scaling
- Splitting Data into Train and Test

PREDICTIVE MODELING

To predict the rate of chronic kidney disease (CKD) among adults, a series of linear regression and ensemble methods were evaluated (see table below). To assess generalizability, the dataset was split into a training and test set. The adjusted R 2 value was calculated to determine the amount of variance in CKD explained by the model while taking the number of features into account. Mean squared error and the corresponding root mean squared error were calculated to measure the predictive accuracy and gain insight into the standard deviation of each model.

	Adjusted R^2	Mean Squared Error	Root Mean Squared Error
Stochastic Gradient Boosting	0.8394	0.1033	0.3214
Extreme Gradient Boosting	0.8377	0.1037	0.3220
Bayesian Ridge Regression	0.8153	0.1180	0.3435
Ridge Regression	0.8147	0.1184	0.3441
Ordinary Least Squares	0.8145	0.1185	0.3442
Random Forest	0.8093	0.1217	0.3489
AdaBoost	0.5994	0.2558	0.5058

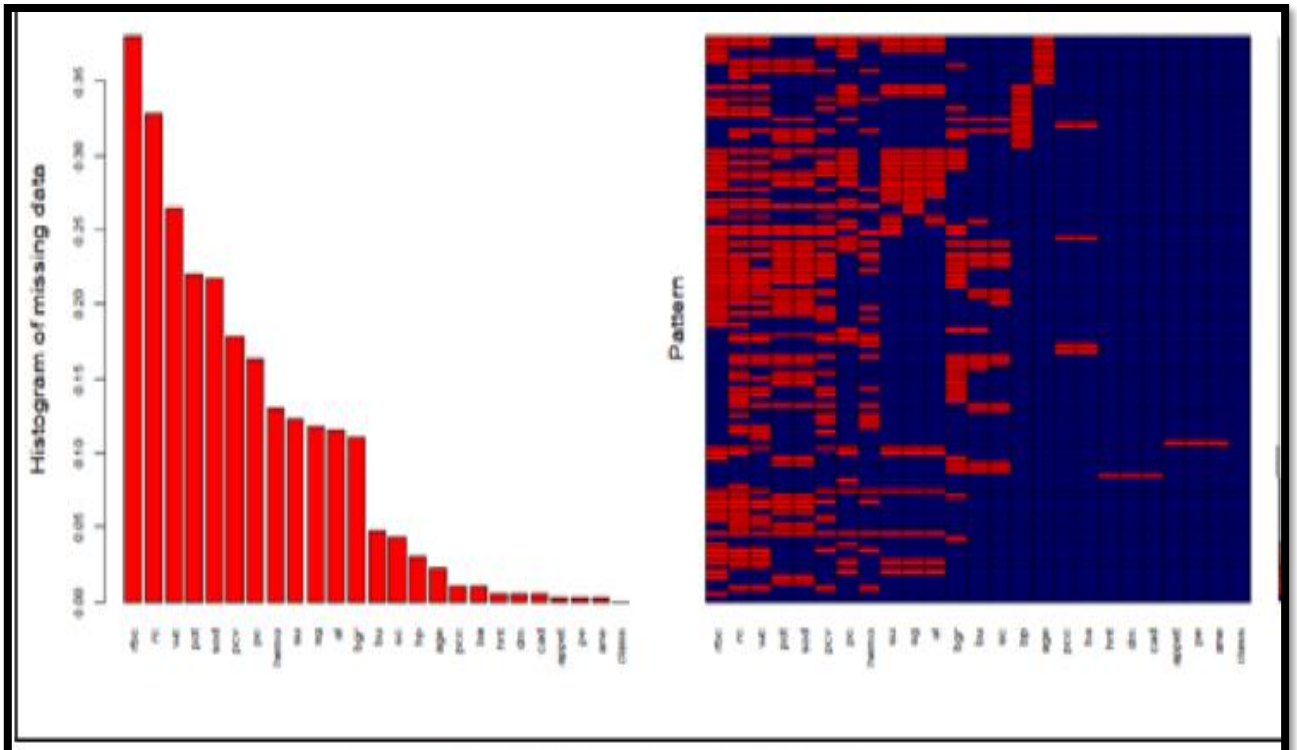


Figure 3: Numerical Predictors and CKD Diagnosis

Another way to deal with these missing values is to impute those.

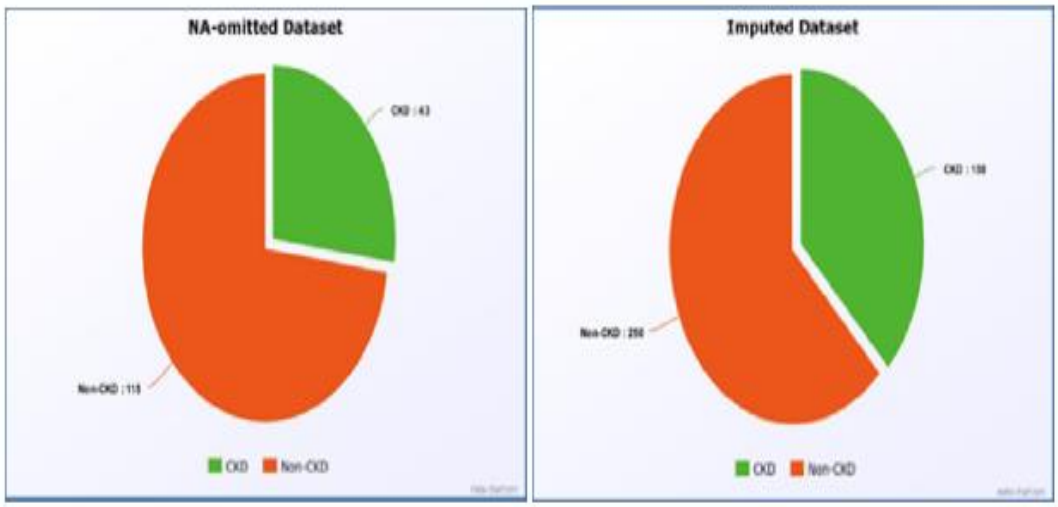


Figure 3:

Numerical Predictors and CKD Diagnosis

Statistical Modelling

Assessment of Current Diagnosis Technique in Literature

As mentioned before, Albuminuria (Al) is a pathological condition in which the protein, albumin, is abnormally present in urine. Persistent Albuminuria suggests some level of kidney damage. Another important factor to measure the severity of CKD is GFR which is estimated by Serum Ceratitine (sc) in combination of age, sex, ethnic origin and body size. To replicate the findings for the current practice, a logistic regression was fit with predictors in Model 1 as just Albuminuria and in Model 2 as serum ceratitine and age where age is the on additional information present in dataset. A low correlation of 0.19 between the two predictors is observed and according to the model, serum creatitine is statistically significant (p-value ¡0.001) but not age, i.e. for one unit increase in serum creatinine the log odds of being CKD increases by 8.81 taking into age into consideration. Hence, the model for the current dataset confirms the current diagnosis technique in literature. Serum creatitine is a good marker for the diagnosis CKD. Table 1 is ummarizes the regression output.

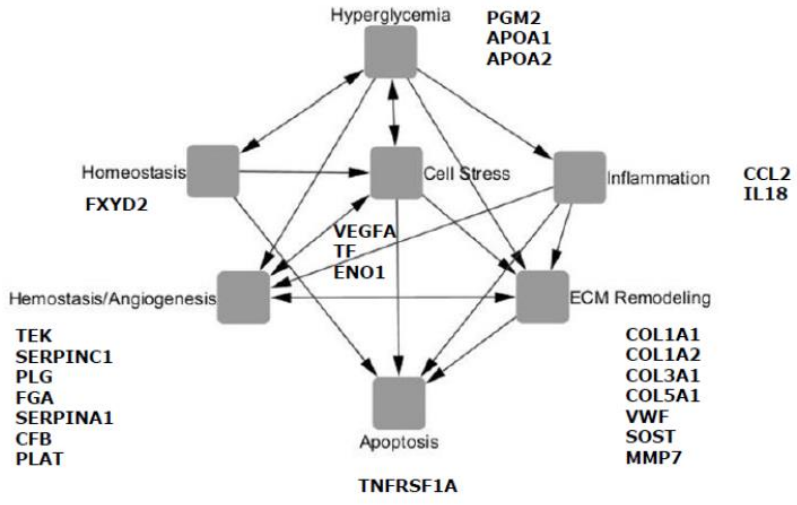


Figure: top-level mechanism perspective on early stage CKD, and candidate biomarkers assigned to specific mechanisms.

Model with Numerical Predictors

When fitting a logistic regression model considering all nominal and numerical predictors, the algorithm does not converge due to perfect separation where the response separates a combination of categorical predictor variables completely. Excluding the categorical variables, the algorithm for the logistic regression model still experience the convergence problem due to high correlation among predictors. An initial heuristic approach is to select subset of six predictors that have low correlation among them. Figure 6 shows the correlation between those selected predictors, age, blood pressure, blood glucose random, sodium, potassium and white blood cell count.

Model Selection

As there is no 'best' model and the model with six predictors in Section 4.2 is chosen by randomly selecting one of the predictors among highly correlated ones, a more methodological way to assess the explanatory and predictive power of predictors is to conduct model selection. Three different model selection techniques are compared. These are stepwise regression, lasso and elastic net.

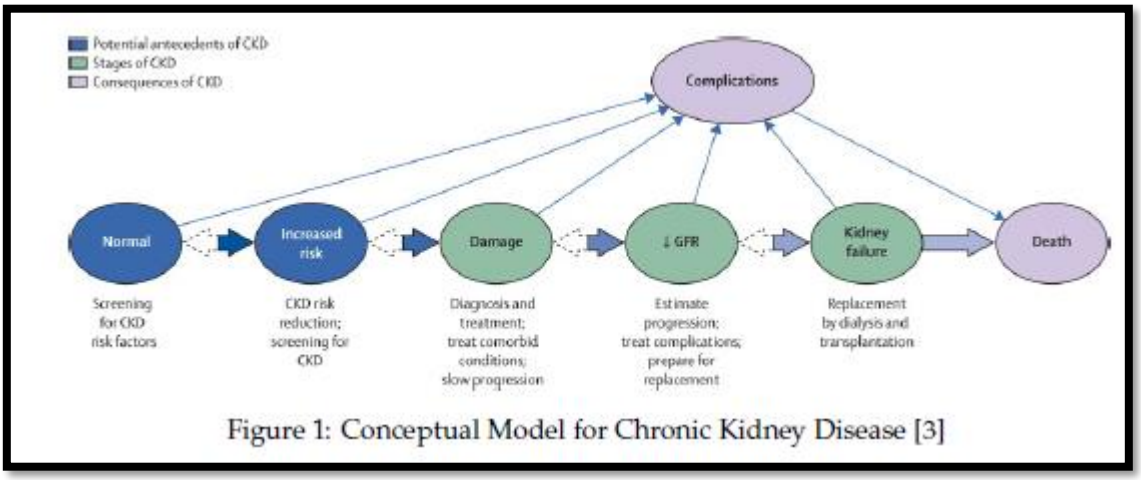


Figure 1: Conceptual Model for Chronic Kidney Disease [3]

Functional components of the project

Data Mining Techniques

1. Support Vector Machines

Support Vector Machines (SVM) is a powerful, state-of-the-art algorithm based on linear and nonlinear regression. Oracle Data Mining implements SVM for binary and multiclass classification. The advantage of the SVM is that, by use of the so-called ‘‘kernel trick’’, the distance between a molecule and the hyper plane can be calculated in a transformed (nonlinear) feature space, lacking of the explicit transformation of the original descriptors. The radial basis function kernel (Gaussian kernel) which is the most commonly used was applied to this study.

2. K-nearest neighbor Classification

In pattern recognition, the K-Nearest Neighbor algorithm (K-NN) is a non-parametric method used for classification and regression. In both cases, the input consists of the K closest training examples in the feature space. K-NN is a type of instance-based learning. In KNN Classification, the output is a class membership. Classification is done by a majority vote of neighbours. If K = 1, then the class is single nearest neighbor. In a common weighting scheme, individual neighbour is assigned to a weight of 1/d if d is the distance to the neighbour. The shortest distance between any two neighbours is always a straight line and the distance is known as Euclidean distance [7]. The limitation of the K-NN algorithm is it’s sensitive to the local configuration of the data. The process of transforming the input data to a set of features is known as Feature extraction. In Feature space, extraction is taken place on raw data before applying K-NN algorithm. The steps involved in a K-NN algorithm:

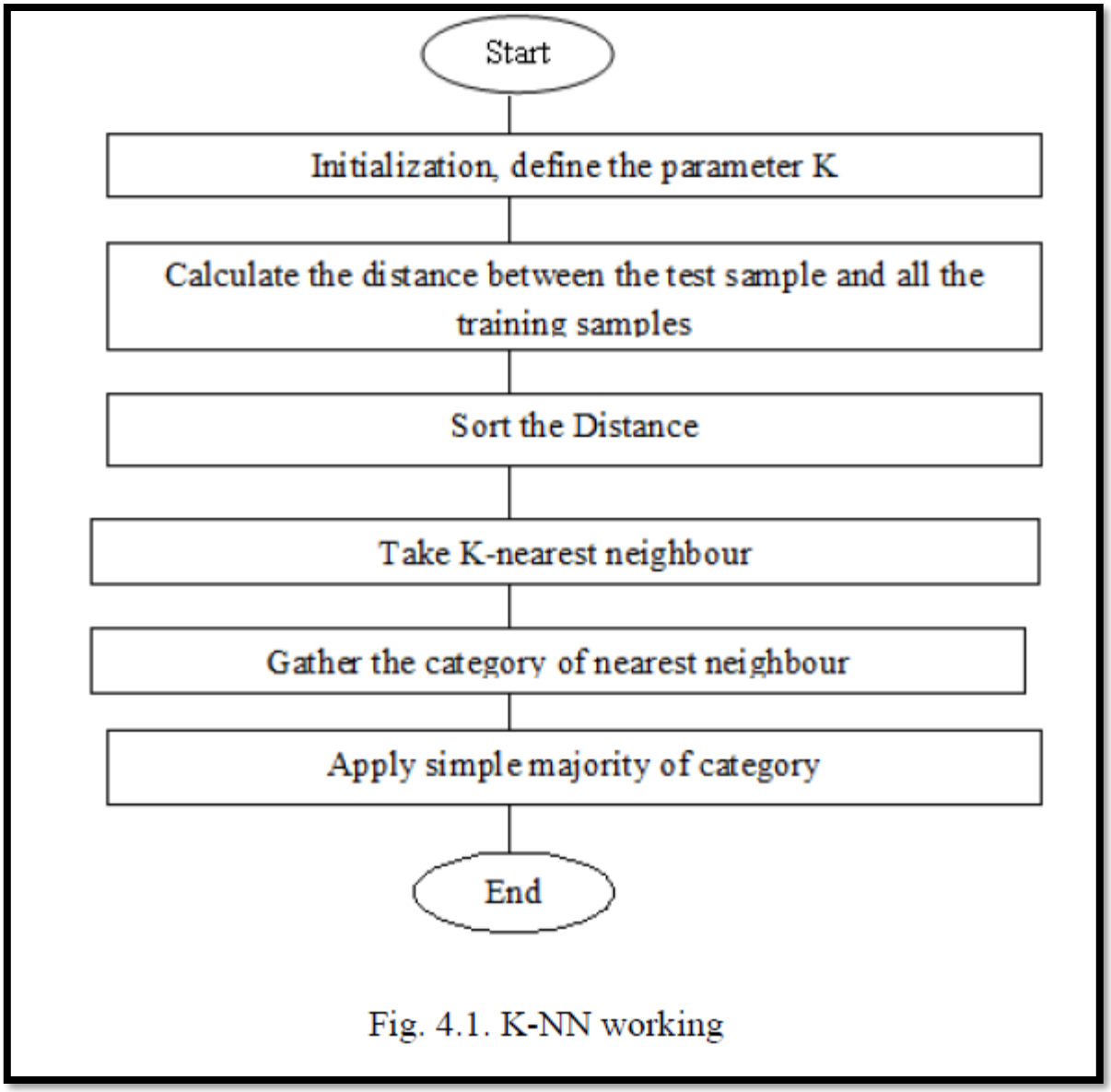
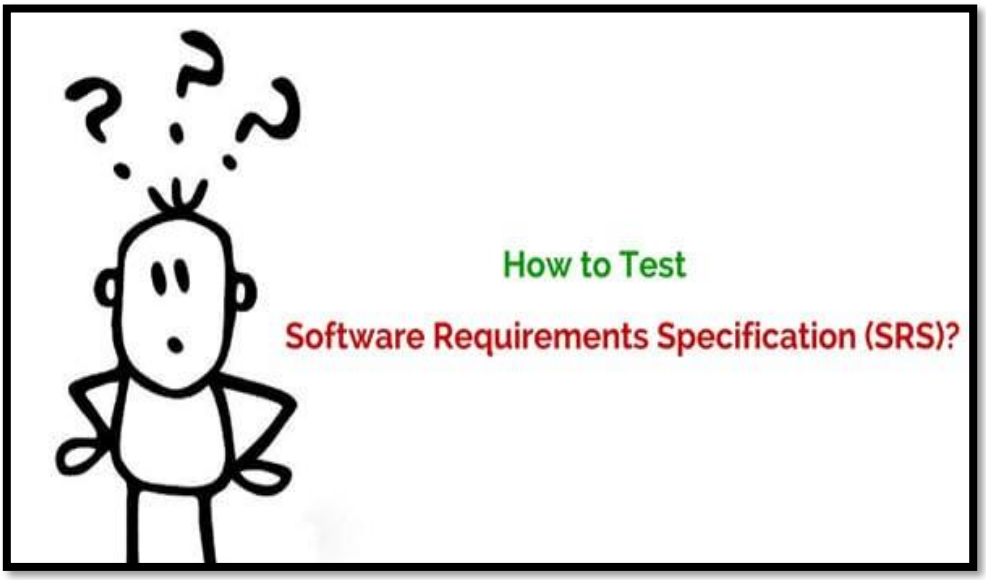


Fig. 4.1. K-NN working

SOFTWARE REQUIREMENT & SPECIFICATION



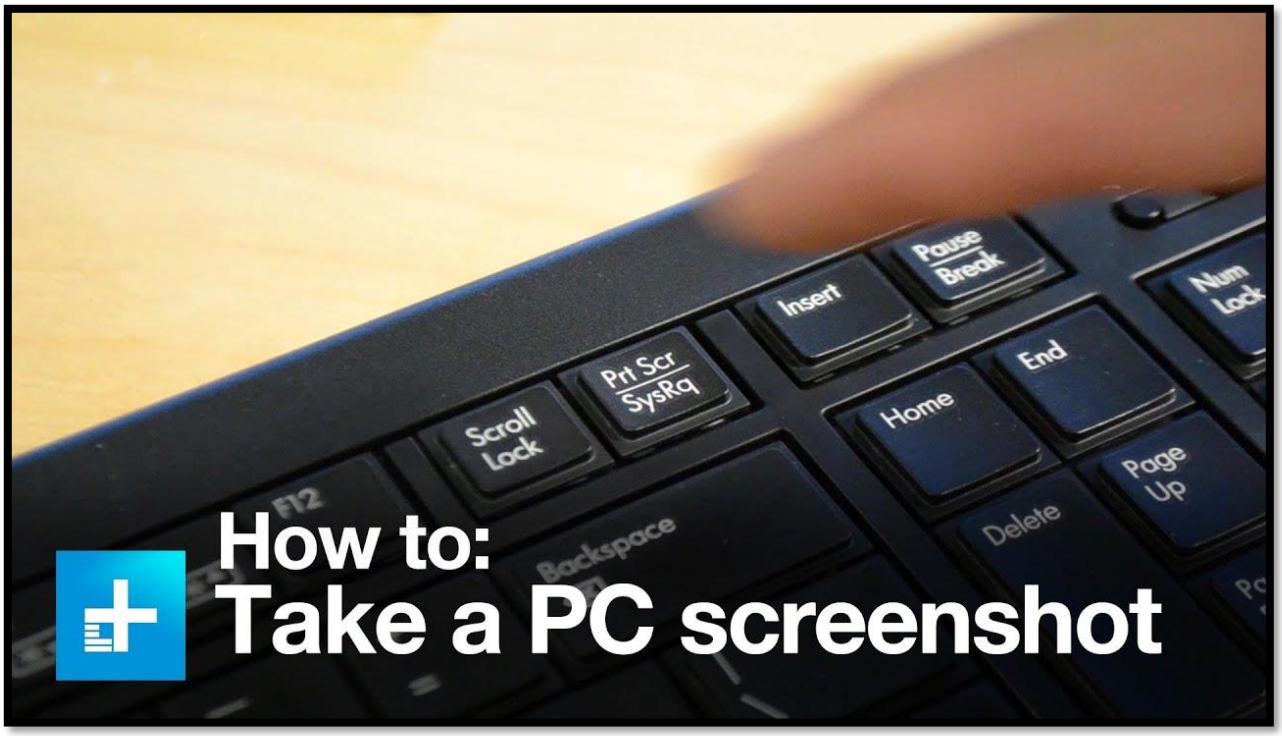
Software, tools & technique Required

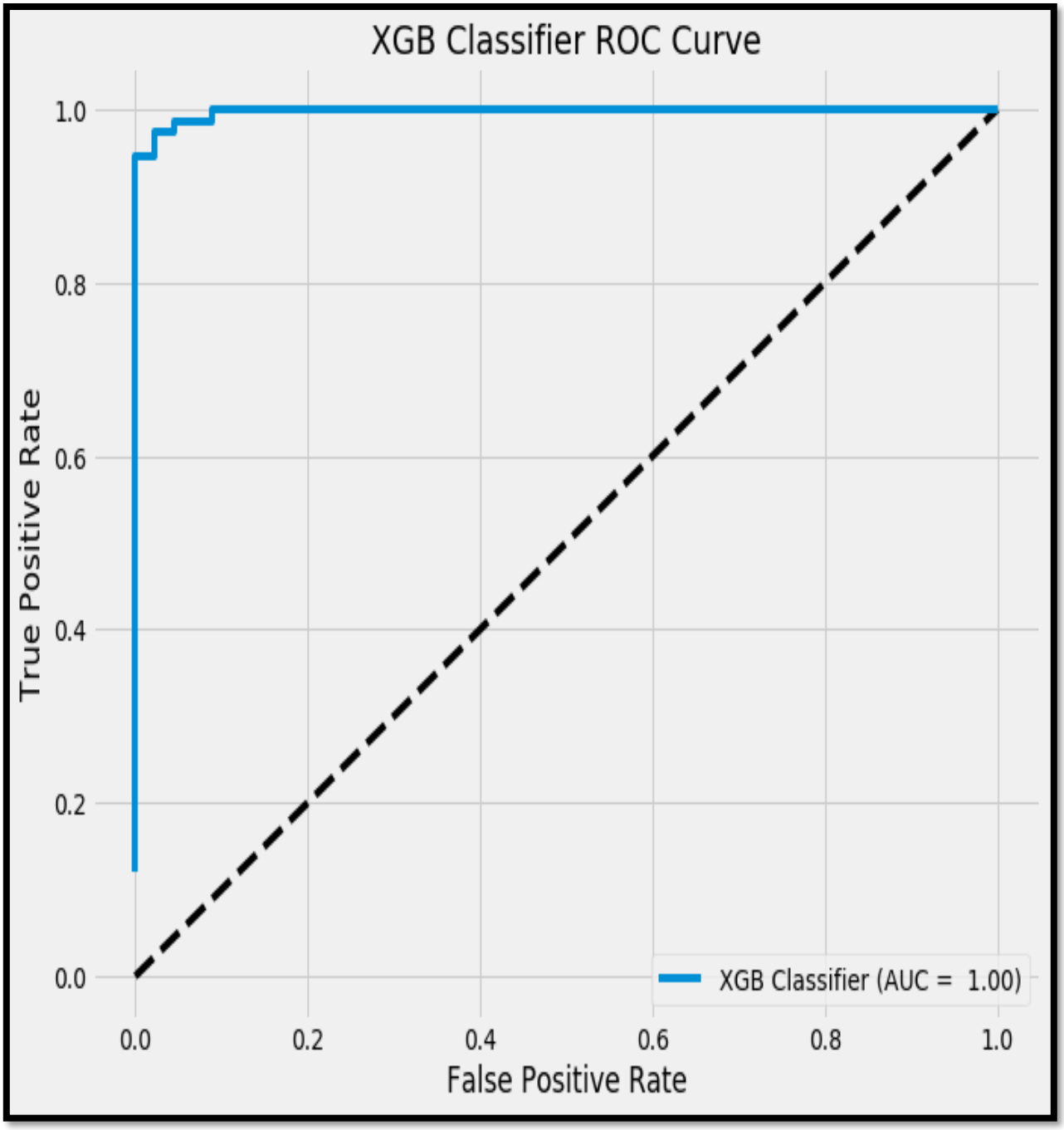
The project is implemented in Python programming language.

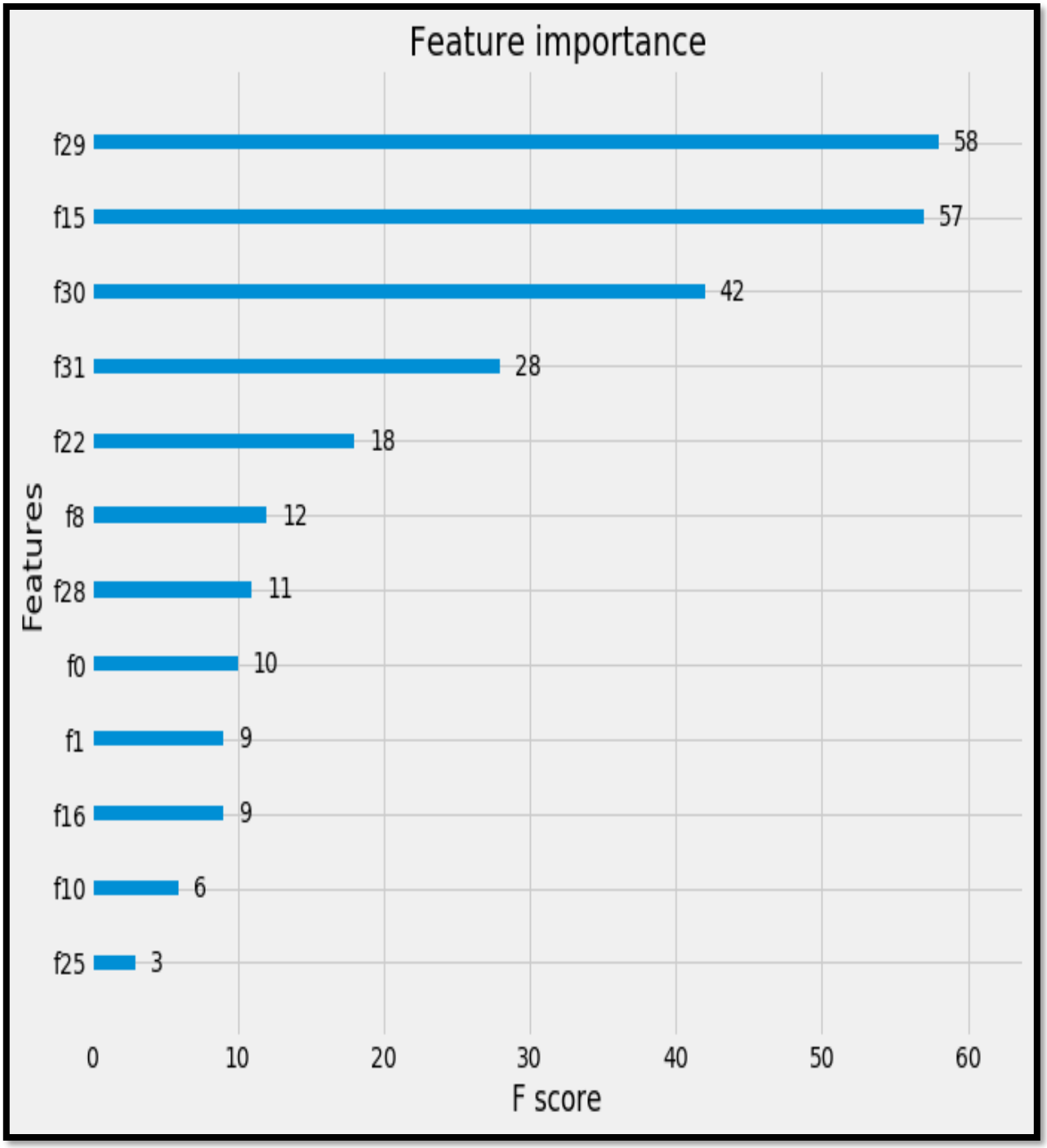
- Python
- Python Web Frame Works
- Python for Data Analysis
- Python for Data Visualization
- Data Pre-processing Techniques
- Machine Learning
- Classification Algorithms

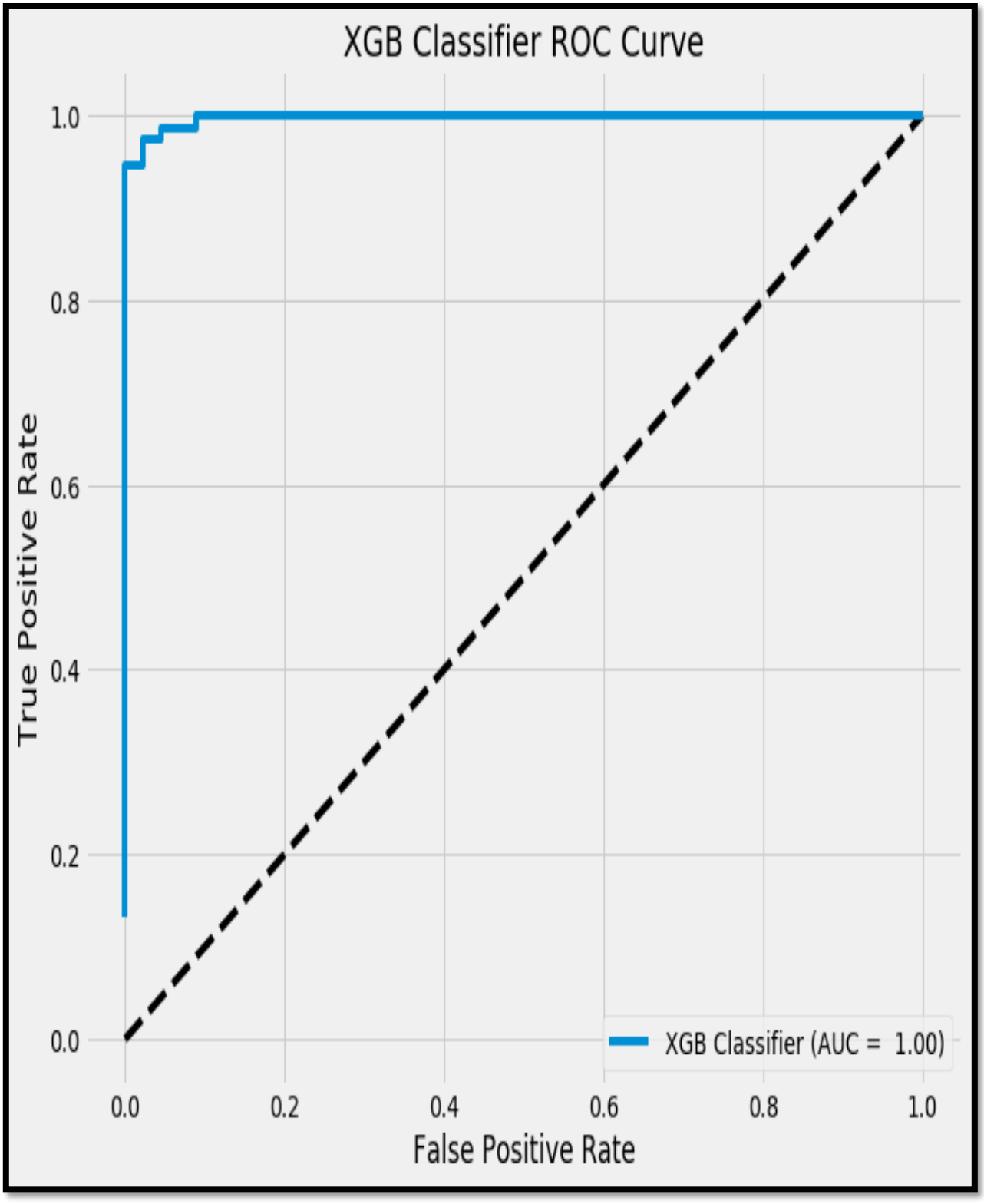
OUTPUT SCREENS

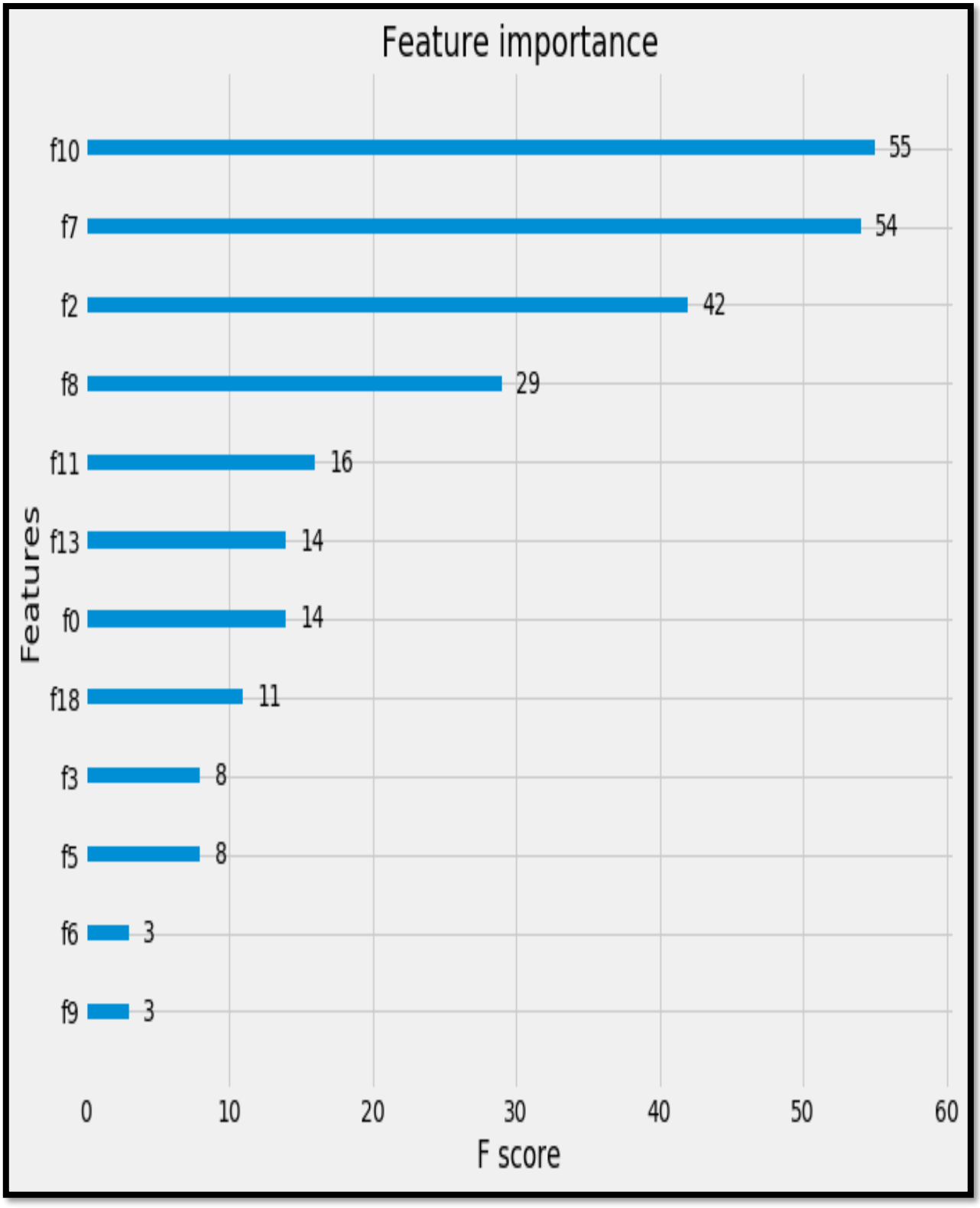
SCREENSHOTS




XGB Classifier ROC Curve
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
XGB Classifier (AUC = 1.00)


Feature importance
f29
f15
f30
f31
f22
f8
f28
f0
f1
f16
f10
f25
58
57
42
28
18
12
11
10
9
9
6
3
Features
0
10
20
30
40
50
60
F score


XGB Classifier ROC Curve
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
XGB Classifier (AUC = 1.00)


Feature importance
f10
f7
f2
f8
f11
f13
f0
f18
f3
f5
f6
f9
55
54
42
29
16
14
14
11
8
8
3
3
Features
0
10
20
30
40
50
60
F score

RESULT AND DISCUSSION

The machine learning methods described are trained to predict the chronic kidney disease. Two classifier methods are used in this decision tree and naive bayes . The experiments are constructed on R tool. In this work , the performance is measured by sensitivity, specificity and accuracy described as follows.
Accuracy (ACC) is the overall success rate of the classifier defined as

ACC = (TP + TN) / (TP + FP + TN + FN)

Sensitivity or the true positive rate (TPR) which is defined as the fraction of positive instances predicted correctly by the model defined as

Sensitivity = TP / (TP + FN).

Specificity is the true negative rate (TNR) which is defined as the fraction of negative instances predicted correctly by the model defined as

Specificity = TN / (FP + TN).

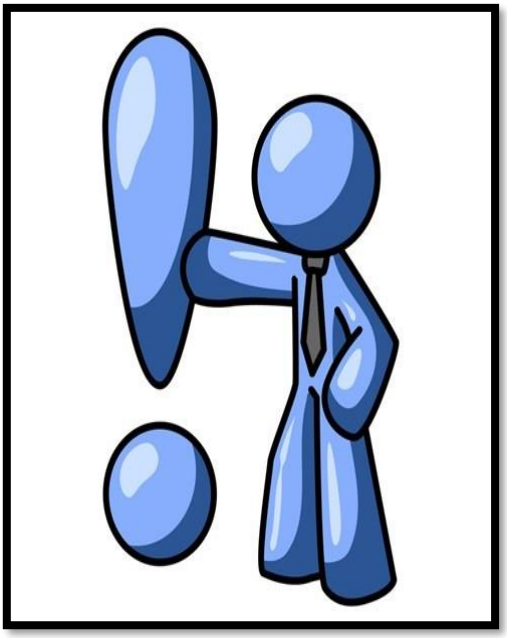
Where
TP - the number of true positives.
TN - the number of true negatives.
FP - the number of false positives.
FN - the number of false negatives.

With the help of True Positive (TP) and True Negative (TN) the performance of the classifications model is evaluated. The machine learning techniques used are trained and tested separately in this work. The 10-fold cross validation is used to train and test the machine learning models in this work and the average results.

The decision tree method and naive bayes method the accuracy of decision tree method is relatively higher than the naive bayes method. The decision tree method can be adopted since it has the accuracy of **99.25%** in prediction of chronic kidney disease.

Techniques used	Accuracy	Sensitivity	Specificity
Decision Tree	99.25%	99.20%	99.33%
Naive Bayes	98.75%	98%	98.75%

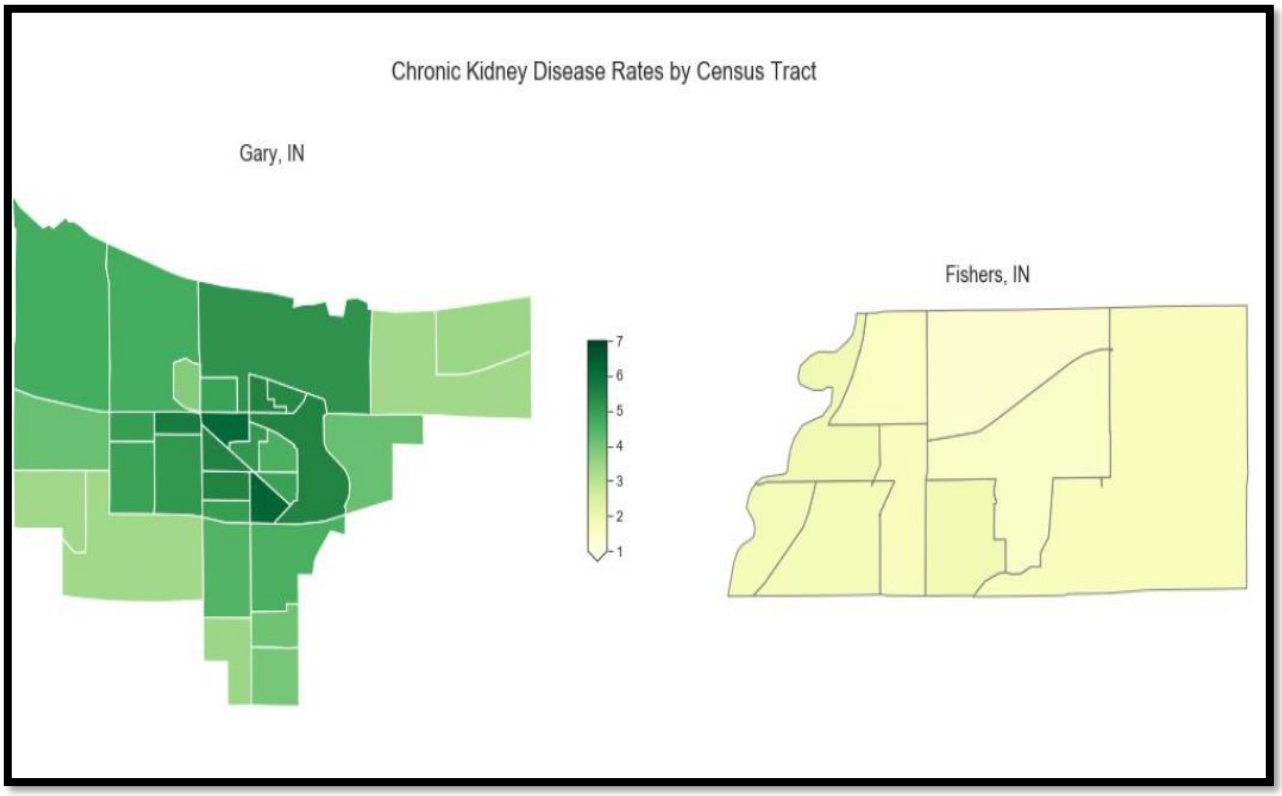
CONCLUSION & SCOPE FOR FUTRURE DEVLOPMENT



CONCLUSION

As we have already seen the applications of data mining and machine learning in medical sector. In this paper, a new decision support system is implemented for prediction of CKD. Although the classifiers worked efficiently in prediction of other diseases also. In this paper, Chronic Kidney Disease is predicted using two different classifiers and a comparative study of their performance is done. From the analysis we found that, out of two classifiers SVM and KNN, KNN classifier performed better than the other. The rate of prediction of CKD is improved.

As we can see in the map generated below, which depicts the rates of CKD at a tract-level within two cities in northern Indiana, this information can allow us to focus limited resources not only on specific counties or cities, but to also narrow campaigns to very specific areas within cities. This allows campaigns to be more focused and cost effective. By looking at feature importance and correlated factors, campaigns can focus on those features (such as employment, ambulatory disability, preventative care, and living alone) that, if affected, would have the most impact on decreasing rates of chronic kidney disease.



SCOPE FOR FURTHER DEVELOPMENT

There are other possible evolutionary techniques that may be used to improve results of the proposed classifiers. In this paper, SVM and KNN are applied to detect CKD. We can also evaluate and compare the performance of the used classifiers with other existing classifiers. CKD early detection helps in timely treatment of the patients suffering from the disease and also to avoid the disease from getting worse. Early prediction of the disease and timely treatment are the need for medical sector. New classifiers can be used and their performance can be evaluated to find better solutions of the objective function in future work.

SCOPE OF ENHANCEMENT

Some people suffering with severe cases of COVID-19 are showing signs of kidney damage, even those who had no underlying kidney problems before they were infected with the coronavirus. Early reports say that up to 30% of patients hospitalized with COVID-19 in China and New York developed moderate or severe kidney injury. Reports from doctors in New York are saying the percentage could be higher.

Signs of kidney problems in patients with COVID-19 include high levels of protein in the urine and abnormal blood work.

The kidney damage is, in some cases, severe enough to require dialysis. Some hospitals experiencing surges of patients who are very ill with COVID-19 have reported they are running short on the machines and sterile fluids needed to perform these kidney procedures.

“Many patients with severe COVID-19 are those with co-existing, chronic conditions, including high blood pressure and diabetes. Both of these increase the risk of kidney disease,” Sperati says.

But Sperati and other doctors are also seeing kidney damage in people who did not have kidney problems before they got infected with the virus.

BIBLIOGRAPHY

For Python installation

https://www.python.org/downloads/

For Anaconda installation

https://www.anaconda.com/products/individual

For Jupyter installation

https://jupyter.org/

For Flask installation

https://pypi.org/project/Flask/

REFERENCES

1. National Kidney Foundation. About Chronic Kidney Disease. National Kidney Foundation A to Z Health Guide.

2. National Kidney Foundation. End Stage Renal Disease in the United States. National Kidney Foundation

3. Soltanpour Gharibdousti, Maryam & Azimi, Kamran & Hathikal, Saraswathi & H Won, Dae. (2017). Prediction of Chronic Kidney Disease Using Data Mining Techniques.

4.Tangri N, Stevens LA, Griffith J, et al. A Predictive Model for Progression of Chronic Kidney Disease to Kidney Failure.

5. Centers for Disease Control and Prevention. About the CKD Initiative.
.

6. Centers for Disease Control and Prevention. Age-adjusted prevalence of CKD Stages 1-4 by Gender

Reference websites

- https://www.kidney.org/atoz/content/about-chronic-kidney-disease . Last accessed October 30, 2018
- https://www.kidney.org/news/newsroom/factsheets/End-Stage-Renal-Disease-in-the-US . Lastaccessed October 30, 2018.
- https://www.cdc.gov/kidneydisease/about-the-ckd-initiative.html . Last accessed November 1, 2018
- . Chronic Kidney Disease (CKD) Surveillance Project website. https://nccd.cdc.gov . Last accessed November 6, 2018.
- www.stackoverflow.com
- www.youtube.com
- https://towardsdatascience.com/machine-learning/home
- https://www.kaggle.com/

Reference Books

- **The Elements of Statistical Learning: Data Mining, Inference, and Prediction**
  -- By: Trevor Hastie, Robert Tibshirani, and Jerome Friedman

- **Applied Predictive Modeling -- By:** Max Kuhn and Kjell Johnson
- **Hands-On Machine Learning with Scikit-Learn and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems -- By:** Aurélien Géron
- **Python Machine Learning -- By:** Sebastian Raschka and Vahid Mirjalili
- Schaum's Out Line: Programming with Python

APPENDIX

Attribute Name	Abbreviation	Unit	Category
Age	age	years	Numerical
Blood Pressure	bp	mm/Hg	Numerical
Specific Gravity	sg	(1.005,1.010,1.015,1.020,1.025)	Nominal
Albumin	al	(0,1,2,3,4,5)	Nominal
Sugar	su	(0,1,2,3,4,5)	Nominal
Red Blood Cells	rbc	(normal,abnormal)	Nominal
Pus Cell	pc	(normal,abnormal)	Nominal
Pus Cell Clumps	pcc	(present,notpresent)	Nominal
Bacteria	ba	(present,notpresent)	Nominal
Blood Glucose Random	bgr	mgs/dl	Numerical
Blood Urea	bu	mgs/dl	Numerical
Serum Creatinine	sc	mgs/dl	Numerical
Sodium	sod	mEq/L	Numerical
Potassium	pot	mEq/L	Numerical
Hemoglobin	hemo	gms	Numerical
Packed Cell Volume hemo	pcv	-	Numerical
White Blood Cell Count	wc	cells/cumm	Numerical
Red Blood Cell Count	rc	millions/cmm	Numerical
Hypertension	hnt	(yes/no)	Nominal
Diabetes Mellitus	dm	(yes/no)	Nominal
Coronary Artery Disease	cad	(yes/no)	Nominal
Appetite	appet	(yes/no)	Nominal
Pedal Edema	pe	(yes/no)	Nominal
Anemia	ane	(yes/no)	Nominal
Class	class	(ckd,notckd)	Nominal

Table 10: Data Description